ARMY RESEARCH LABORATORY

Exploratory Data Analytics for Information Discovery in a Network Structure

by Andrew M. Neiderer

ARL-TN-462 **November 2011**

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5067

ARL-TN-462 **November 2011**

Exploratory Data Analytics for Information Discovery in a Network Structure

Andrew M. Neiderer
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)*
November 2011

2. REPORT TYPE
Final

3. DATES COVERED (From - To)
October 2010–August 2011

4. TITLE AND SUBTITLE
Exploratory Data Analytics for Information Discovery in a Network Structure

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Andrew M. Neiderer

5d. PROJECT NUMBER
1TEDUC

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Laboratory
ATTN: RDRL-CII-C
Aberdeen Proving Ground, MD 21005-5067

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TN-462

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This report presents an analytic strategy for visual exploration of multidimensional data. Node position in a network structure is determined by projecting from the high-dimensional data (HDD) space to a low-dimensional latent space. Clustering of node position vectors may result for making inferences. Dimensionality reduction by feature extraction of HDD for visualization is performed using a parametric Student's t-distribution for stochastic neighbor embedding (t-SNE). The resultant t-SNE network of nodes for a Euclidean space can now be examined using visual analytics technology—navigation/interaction within the visualization of the data. Scene content is described using the Extensible 3-D (X3D) graphics application programming interface. The immersive profile of an X3D scene allows for navigation within the data for possible information discovery. Such an approach may provide for a better understanding of data and facilitate analytical reasoning that would otherwise be difficult in an exclusively textual context.

15. SUBJECT TERMS
dimensionality reduction, parametric t-distributed stochastic neighbor embedding, visual analytics, X3D, network structure

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Andrew M. Neiderer
					19b. TELEPHONE NUMBER (*Include area code*)
Unclassified	Unclassified	Unclassified	UU	20	410-278-3203

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

1. Introduction

A challenge in analyzing terrorist threats is separating the relevant information that is often buried within a massive amount of other data. This relevant (or supervised) data must usually be further reduced, especially when humans are involved in an interpretation. Even identifying simple relationships from a text extraction of data can be a challenge and is usually easier and more quickly comprehended when presented graphically. Therefore, transformation of all that is known about the data to a reduced set is welcomed. Then, allowing for exploration (navigation and interaction) in two-dimensional/three-dimensional (2-D/3-D) data prior to an arbitrary projection may result in information discovery.

The U.S. Army Research Laboratory (ARL) is addressing this complex topic by developing software that includes dimensionality reduction (DR) for data analytics (DA) and subsequent application of visual analytics (VA) technology to take advantage of the broad eye/brain pathway. This human combination is amazingly efficient at analyzing and interpreting massive amounts of data when presented in an effective visual format—more of the brain is devoted to visual processing than to any other sense. Lee and Verleysen (*1*) state that humans try to understand high-dimensional structures in the same way as 2-D/3-D objects. When the dimensionality is more than three (e.g., 16 features to be represented by a single pixel), it is difficult and often confusing to try to perceive similarities/dissimilarities in the data. The following application feature extraction is done using a "think globally, fit locally" approach as opposed to a simple selection of features in the data. It is a nonlinear DR (NLDR) approximation that preserves topology when projecting from high-dimensional data (HDD) space to a 2-D latent space.

The next section discusses an algorithm being considered for NLDR—a parametric Student's t-distributed stochastic neighbor embedding (t-SNE) (*2*) for a rapid mapping of feature data from HDD space (d) to latent space (X). The t-SNE preserves the topology (*3*)[*] of the data after an extraction, which may be important since dependencies could exist between nodes. This intrinsic property is not altered when projecting from d to X; deformation, twisting, and/or stretching (intrusions) are allowed but no tearing (extrusion). For example, in 2-D Euclidean space, a circle is topologically equivalent to an ellipse, but when you tear (or cut) it, you lose the topological structure, and one now has a random line segment.

Section 3 describes the VA capability for interaction with data. A scene is described using the Extensible 3-D (X3D)[†] application programming interface for the data. The X3D is an

[*]In topology, the concern is not the representation of an object or structure in space but connectivity.

[†]Note that the functionality of an X3D node and its attributes are described at http://www.web3d.org/x3d/content/x3dTooltips.

International Standards Organization (ISO) specification that allows for real-time, interactive manipulation of data in a scene possibly distributed across the Web. In late 2010, X3D nodes were tightly coupled within the hyper-text markup language (HTML) document object model tree of Web browsers, such as Internet Explorer.* A European Computer Manufacturers Association Scripting (ECMAScript)-language access to scene content for interaction is done through an X3D <Script> node.

An example is given throughout the report for navigation within a visualization of a network of nodes. The parametric t-SNE (*4*) is programmed in MATLAB (*5*). The VA capability was done for the Xj3D standalone browser Xj3D 2_M1_DEV_2008-05-08† developed at Yumetech, Inc. The VA program is written for stereo viewing in an immersive profile.

This initial research has not yet been finalized. The VA work has been finalized, as demonstrated for navigation within a visualization of a network of nodes. Although the parametric t-SNE has been successfully used with the MNIST database of handwritten digits (*6*), it has yet to be used with terrorist data.

2. Dimensionality Reduction: The Data Analytics for Visual Analytics

Visualization of any underlying structure that may exist for real-world HDD involves a projection to a plane in 2-D space. DR aims at an extraction of features (as opposed to simple feature selection) by eliminating any redundancy that may exist. However, preserving structure or dependencies within the data is important so that there is no loss of information when re-embedding the "true" manifold from d to one in this lower dimension, or the projected manifold must remain representative of the actual data and topological properties not altered.

DR tries to exploit the typically lower intrinsic dimension (P) of the real-world data, i.e., for $P<d$. P is the minimum number of parameters needed to account for observed properties of the data and reveals the presence of topological structure in the data. Ideally, the reduced dimension (D) will correspond to P. When $P \le D$ where D is also the dimension of the embedding space, the data lie in a well-defined space. The most common way to estimate P is by computing the number of latent variables.

A leading researcher in data visualization, John A. Lee, describes a manifold as a topological space that is locally Euclidean but may be globally curved (*7*). He also states in his book that a topological object is formally defined as a topological space. For example, the Earth is spherical in shape but looks flat to the human eye. Topology abstracts the intrinsic connectivity of an

*Internet Explorer is a registered trademark of Microsoft Corporation.

†The viewer can be found at http://www.xj3d.org/snapshots.html. Java-language bindings are also defined for manipulating /viewing scene content but not used here.

object (or structure) but ignores the detailed form. Each point in the original HDD is assumed to lie near or on a manifold and should remain close or on a manifold after re-embedding in R^D, where R is real and $D \leq d$; D is either a 2- or 3-D embedding space that is Euclidean. The embedding space R^2 is the latent space for reduced data. Lee and Verleysen recently stated that DR is a “boiling hot research topic” (*7*). For a linear DR (LDR) such as principal component analysis (PCA) or classical multidimensional scaling (MDS), the metric is based on Euclidean distance between two points and is called distance-preserving. However, LDRs cannot handle complex, nonlinear cases typical of real-world data. Thus, a NLDR approximation (or manifold learning) based on the geodesic distance along the manifold (linear or nonlinear) is used instead of a Euclidean distance for the metric; this approach is topology-preserving. An example comparing the application of an LDR to NLDR for HDD is illustrated in figure 1; a projection from a 3-D embedding space $Y = [y_1\ y_2\ y_3]$ to a 2-D latent space $X = [x_1\ x_2]$ shows the concern.

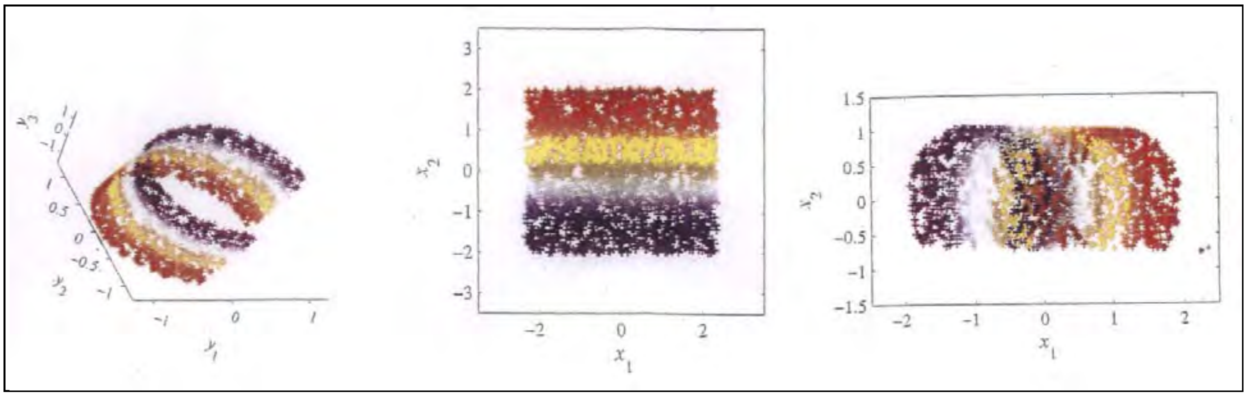

Figure 1. Image showing (a) comparison of an NLDR vs. LDR for a 2-D manifold embedded in a 3-D space Y, (b) using a NLDR, and (c) using a LDR. It should be noted that for a NLDR, the topology is preserved. The figures are taken from Lee and Verleysen (1).

There are many NLDR techniques. Manifold learning has been successfully demonstrated for artificial datasets such as the Swiss roll, where points lie on a spiral-like 2-D manifold embedded in 3-D Euclidean space. NLDRs find this embedding, whereas LDRs fail to do so. NLDRs have been quite successful on artificial datasets but less convincing on natural datasets, where real-world data are typically highly curved. Now, recent research (*8*) suggests that DRs for learning manifolds differ from DRs for data visualization. Both of these concerns (real-world data and data visualization) are being considered.

Additionally, a near real-time capability may be imperative. A parametric t-SNE meets this requirement once training for a HDD space to low-dimensional latent space is completed; in fact, the algorithm is faster than PCA, the quickest of all DR algorithms.

In our application, the parametric t-SNE eliminates redundancy of some 16 features when computing latent variables. Specifically, the features are tribal affiliation, probable origin, observer recognition ID, remote sensed facial imagery ID, remote sensed pulse rate, directly

measured pulse rate, directly sensed GSR, iris pattern ID, facial imagery ID, ID according to fingerprint, Taskera name congruent with claimed name, Taskera name congruent with true ID, probable origin, assumed age, probable ethnicity, and recorded sect.

It should be noted that visualization of data resulting from application LDRs/NLDRs is done in a 2-D latent space. A latent variable is at the origin of observed values but cannot be measured directly. Both LDRs and NLDRs find the number of latent variables, but determination of the actual latent variables themselves, known as latent variable separation (LVS), is beyond the scope of this work (LVS, including discussion of the two more popular approaches, blind source separation and independent component analysis, can be found in the book by Hyvarinen et al. [*9*]). In general, however, it is difficult to tell the meaning of latent variables.

3. The X3D Visual Analytic for Network Exploration

A VA capability provides for interaction with data (*10*). In our case, this is navigation within a visualization of nodes for gaining additional, timely insight to network topology or connectivity. For example, a rotation of the scene in figure 2 about the y-axis results in discovery of C3 hidden by C2 (see figure 3); this relationship would be difficult to identify in a text presentation. Such affine transformation(s) of the data have been defined in Neiderer (*11*).

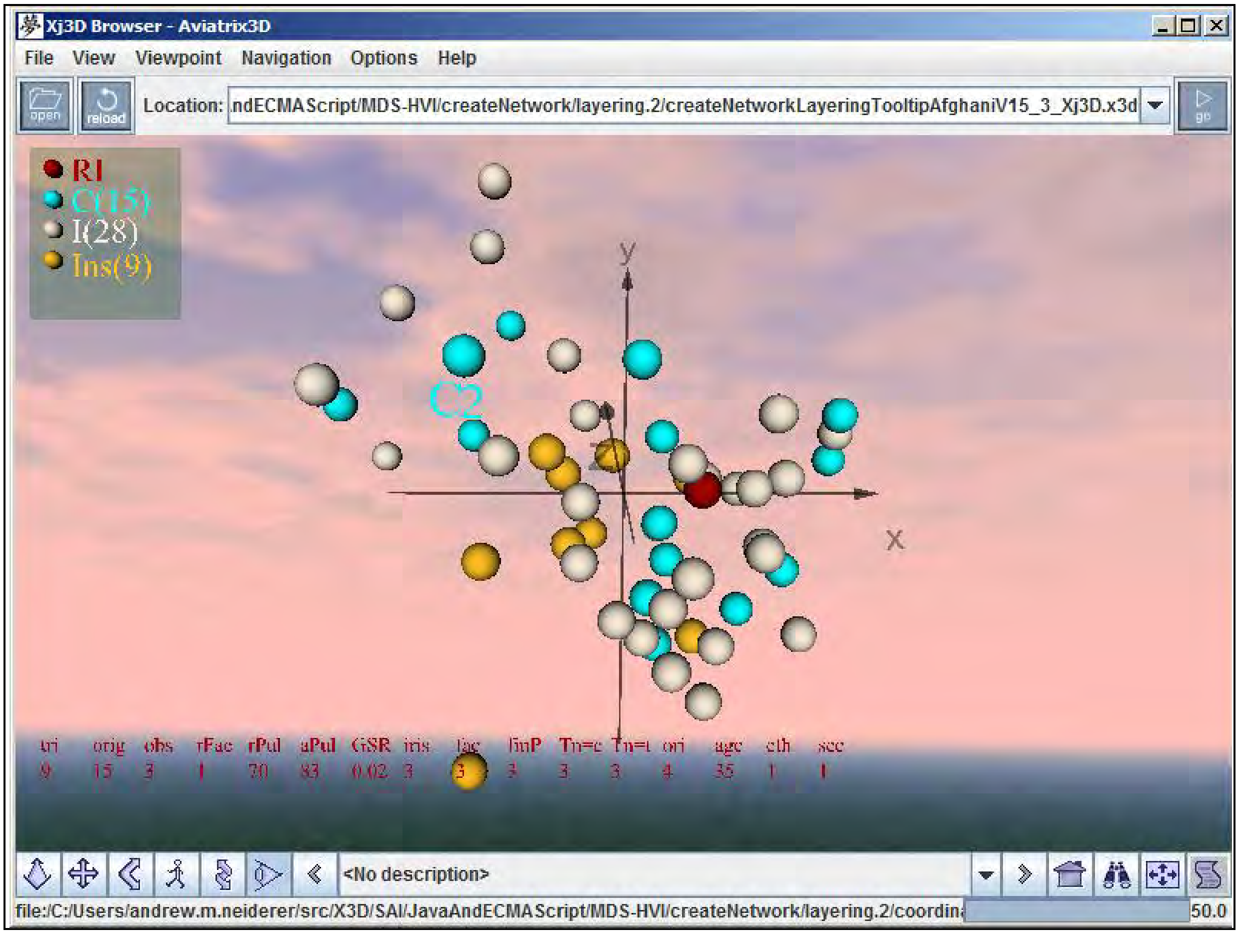

Figure 2. An Xj3D view of a 53-node network. The C2 tooltip is activated when the mouse pointer passes over that node.

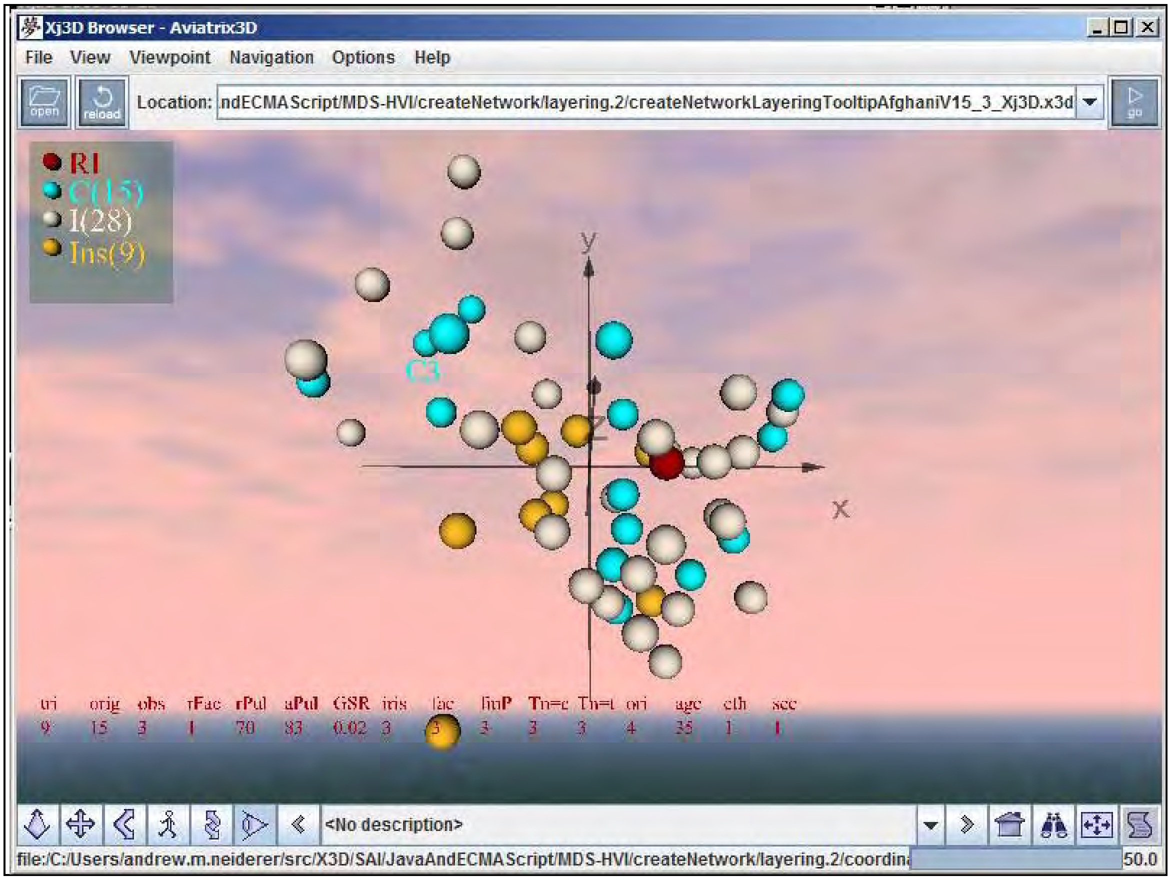

Figure 3. An Xj3D view of a scene from figure 2 that has been rotated about the y-axis for display of C3, which was not visible in the previous figure.

X3D uses an extensible markup language (XML) encoding of data. It continues to expand and be embraced by 3-D computer graphics developers in many different fields. Recently, X3D nodes were tightly coupled with the HTML document object model (DOM) tree of Web browsers (*12*). The result is a seamless integration where X3D programs can be run without changing a single line within an application. For now, however, X3D scenes are displayed in a browser from Yumetech, Inc.

Specifically, X3D nodes, or objects, are viewed in the Xj3D-2_M1_DEV_2008-05-08 browser. Xj3D provides for both Java- and ECMAScript-language bindings to scene content. It is an open-source, standalone browser that supports over 170 X3D primitives, including an unlimited number of prototype definitions. X3D nodes are grouped into a component and components by profiles. The immersive profile for a VA capability is used here. A thorough discussion of these concepts and X3D in general can be found in the book by Brutzman and Daly (*13*).

X3D nodes can be chained together by fields for animation. This is how tooltips are defined in a scene. The <ROUTE> mechanism allows for real-time, interactive manipulation communication with the displayed content.

A detailed description of an entire scene for a network of nodes is given in Neiderer (*11*), as well as the event cascades for animation. Although all details are not repeated here, the scene graph is described and discussed in the next section.

X3D Scene Graph Description

The scene graph (SG) representation for a network of 53 nodes is illustrated in figure 4. The key at the upper left describes the content as follows: an individual of remote inquiry (RI), 15 criminals (C), 28 innocents (I), and 9 insurgents (Ins). The console across the bottom is used for both static and dynamic display of node features—the 16 attributes of the RI are static and can be compared to any node in the scene by "touching." For example, in figure 4, attributes of C1 can be compared to the RI.

Each network node has two branches, both directed acyclic graphs (DAG) of X3D objects—a geometry branch and a text branch. In this way, we keep the geometry in a scene separate from text. The branches are fully described in Neiderer (*11*), and only the figure is repeated here (see figure 5).

That report also discusses the event chaining for fields of X3D nodes defined for tooltip and dynamic display of text. Figure 2 displays the situation where the mouse pointer passes over C2 (criminal 2); the result is a tooltip for quick identification of that node. This can be done for any node in the scene. A second event chain is defined for clicking (or "touching") any node in a scene, and the appropriate text is routed to the console at the bottom of the display (see figure 4). It should be noted that both the key and console have been placed in a layer separate from scene content. This allows for navigation within a scene and independent text display. In this case, text is always displayed left to right in the same location.

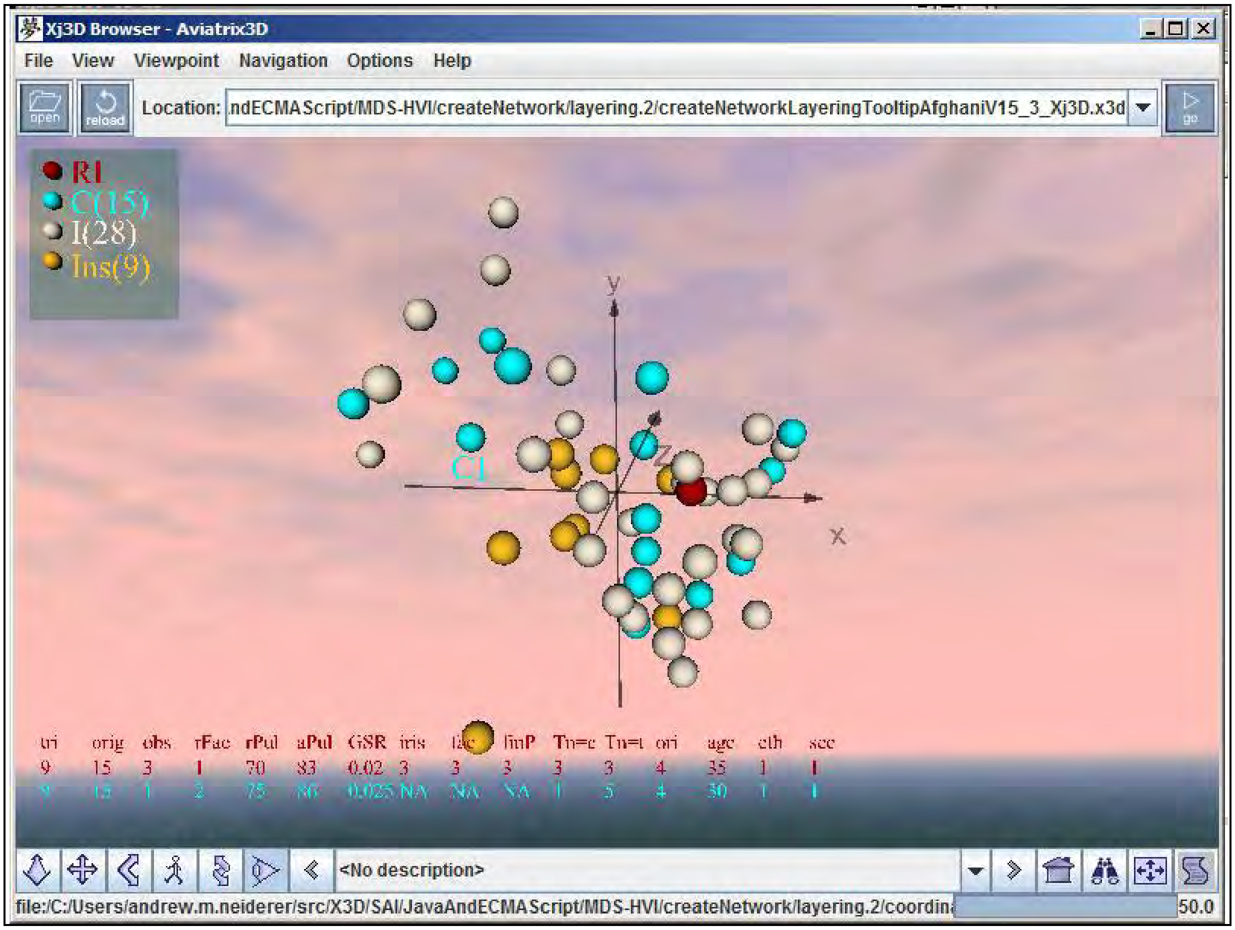

Figure 4. An Xj3D view of a 53-node network with a legend (at left) and a console (bottom). Network node “C1” is touched, resulting in text animation for the node that can be compared to the “RI.”

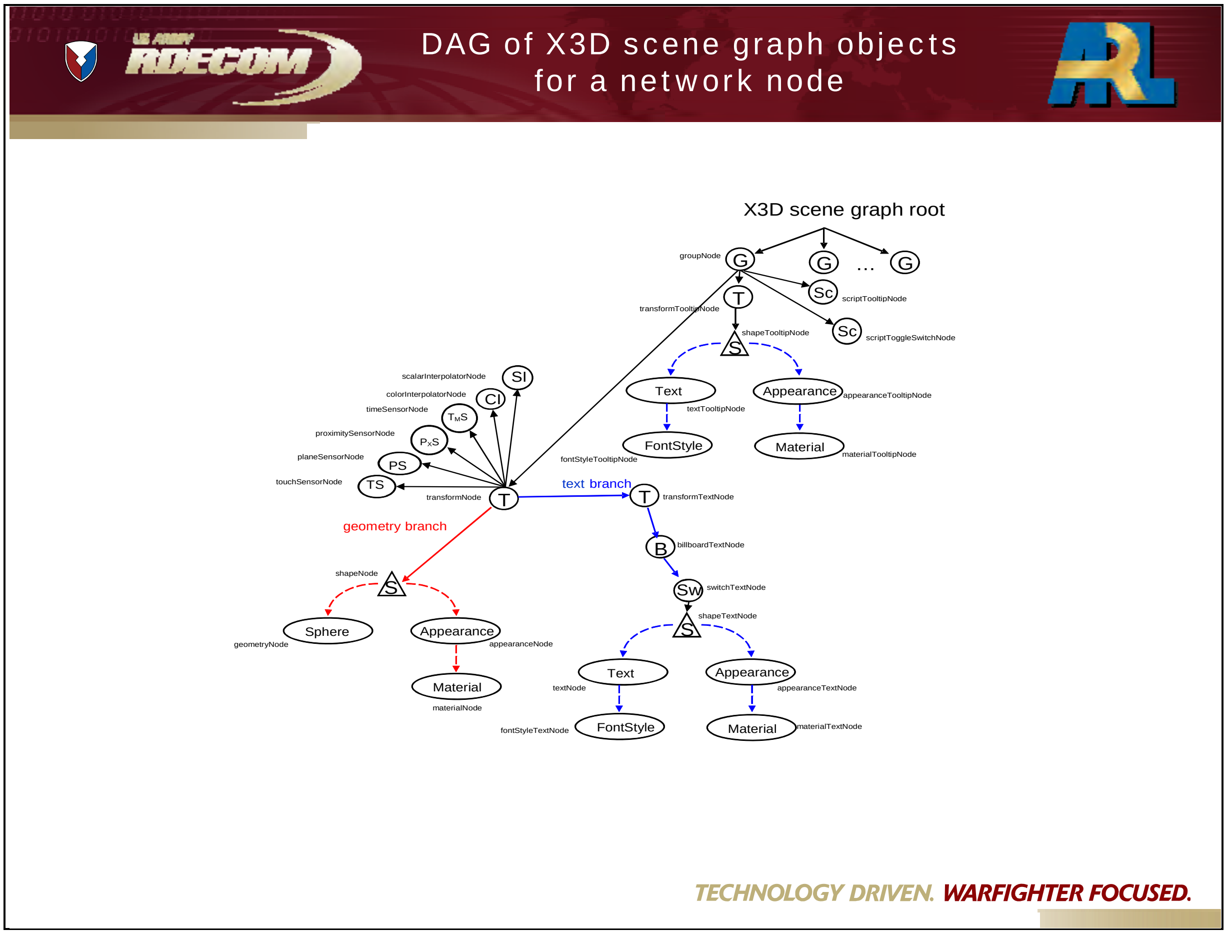

Figure 5. DAG of X3D scene graph objects for a network node.

4. Conclusions and Future Work

Dimensionality reduction for visual analytics is being developed at ARL for exploratory data analysis. VA for a network structure has been completed using the X3D standard application programming interface. Currently, the application of a parametric t-SNE algorithm, which reduces terrorist data to node position vectors of a potential terrorist network, is being considered. This NLDR uses distances along the manifold in HDD space (i.e., geodesic distances) so that a re-embedded manifold is topologically preserved. For VA, the dimensionality of the reduced data is for a 2-D/3-D latent space. Examining the data in this context may result in identification of key relationships. A parametric approach will allow for new data entry and fast evaluation once learned. Future work will expand the number of DR techniques and utilize data extracted from simulated and real intelligence reports.

5. References

1. Lee, J. A.; Verleysen, M. *Nonlinear Dimensionality Reduction*; Springer Science + Business Media: New York, NY, 2007.

2. Van der Maaten, L.; Hinton, G. Visualizing Data Using t-SNE. *J. of Machine Learning Research* **2008**, *9*, 2579–2605.

3. ISO/IEC 19775: 2004 Extensible 3-D Standard. http://www.web3d.org/x3d/specifications /x3d_specification.html (accessed September 2011).

4. t-SNE. http://homepage.tudelft.nl/19j49/t-SNE.html (accessed September 2011).

5. The Mathworks, Inc. http://www.mathworks.com (accessed September 2011).

6. Van der Maaten, L. Feature Extraction from Visual Data. Ph.D. Thesis, Tilburg University, The Netherlands, 2009.

7. Lee, J. A.; Verleysen, M. Unsupervised Dimensionality Reduction: Overview and Recent Advances. *Proceedings of the WCCI 2010 IEEE World Congress on Computational Intelligence*, Barcelona, Spain, 18–23 July 2010.

8. Kaski, S.; Peltonen, J. Dimensionality Reduction for Data Visualization. *IEEE Signal Processing Mag.* **2011**, *100*.

9. Hyvarinen, A.; Karhunen, J.; Oja, E. *Independent Component Analysis*. Wiley-Interscience: New York, NY, 2001.

10. Thomas, J. J.; Cook, K. A. Illuminating the Path. *IEEE Computer Society* **2005**.

11. Neiderer, A. *Network Visualization Using Xj3D*; ARL-MR-754; U.S. Army Research Laboratory: Aberdeen Proving Ground, MD, September 2010.

12. Independent Component Analysis. http://en.wikipedia.org/wiki/Independent_component _analysis (accessed September 2011).

13. Brutzman, D.; Daly, L. *X3D: Extensible 3D Graphics for Web Authors*; Morgan Kaufmann Publishers: San Francisco, CA, 2007.

List of Symbols, Abbreviations, and Acronyms

3-D	three-dimensional
ARL	U.S. Army Research Laboratory
d	data space
D	reduced dimension
DA	data analytics
DAG	directed acyclic graph
DOM	document object model
DR	dimensionality reduction
ECMAScript	European Computer Manufacturers Association Scripting language
HDD	high-dimensional data
HTML	hyper-text markup language
ISO	International Standards Organization
LDR	linear dimensionality reduction
LVS	latent variable separation
MDS	multidimensional scaling
NLDR	nonlinear dimensionality reduction
P	intrinsic dimension
PCA	principal component analysis
SG	scene graph
t-SNE	t-distributed stochastic neighbor embedding
VA	visual analytics
X	latent space
X3D	Extensible three-dimensional language

Xj3D	Extensible three-dimensional language viewer with Java-language bindings
Y	embedding space

NO. OF COPIES	ORGANIZATION
1 (PDF only)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA 8725 JOHN J KINGMAN RD STE 0944 FORT BELVOIR VA 22060-6218
1	DIRECTOR US ARMY RESEARCH LAB IMNE ALC HRR 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	DIRECTOR US ARMY RESEARCH LAB RDRL CIO LL 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	DIRECTOR US ARMY RESEARCH LAB RDRL CIO MT 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	DIRECTOR US ARMY RESEARCH LAB RDRL D 2800 POWDER MILL RD ADELPHI MD 20783-1197

NO. OF COPIES	ORGANIZATION
1	GSC ASSOC INC G S CARSON 2727 XANTHIA CT DENVER CO 80238-2611
1	DIR USARL RDRL CIN A KOTT 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	DIR USARL RDRL CII B BROOME 2800 POWDER MILL RD ADELPHI MD 20783-1197
	ABERDEEN PROVING GROUND
5	DIR USARL RDRL CII C A NEIDERER (4 CPS) M THOMAS